Mathematics Education Research Journal 2006, Vol. 18, No. 2, 77-92

A Case of the Inapplicability of the Rasch Model: Mapping Conceptual Learning

Kaye Stacey and Vicki Steinle
University of Melbourne

The basic theory of Rasch measurement applies to situations where a person has a certain level of a trait being investigated, and this level of ability is what determines (to within a measurement error) how well the person does on each item in a test. This paper responds to frequent suggestions from colleagues that the use of Rasch measurement would be profitable in analysing a set of data on students' understanding of decimal notation. We demonstrate misfit to the Rasch model by showing that item difficulty estimates show important variation by year level, that there is significant deviation from expected score curves, and that success on certain splitter items does not imply a student is more likely to score well on other items. The explanation given is that conceptual learning may not always be able to be measured on a scale, which is an essential feature of the Rasch approach. Instead, students move between categories of interpretations, which do not necessarily provide more correct answers even when they are based on an improved understanding of fundamental principles. In this way, the paper serves to highlight the assumptions built into the Rasch model and to discuss its applicability to describing the progress of learning with various characteristics.

The basic theory of Rasch measurement has been developed to apply fundamental principles of physical measurement to human traits and abilities. To weigh an apple, one finds out whether it is heavier or lighter than a range of calibrated weights and this enables it to be placed appropriately on an agreed numerical scale. Similarly, a Rasch model aims to measure the ability of a person in a domain by testing it against a set of items of calibrated difficulty, so that the person's ability can be placed appropriately on the scale. The basic Rasch model specifies that a person has a certain level of the latent trait being investigated, and that it is only this level of ability and the difficulty of the item that determines (to within measurement error) how well the person performs on an item. If these assumptions hold for an area of learning, it follows that as students learn more, their ability in the area will increase and hence they will be able to answer correctly items of increasing difficulty, whilst retaining the ability to answer the easier items. If students of lower ability answer an item correctly more often than students of higher ability, the item is regarded as suspect and for many committed to Rasch test construction such an item simply would be discarded from the test. To develop a scale, a person's ability and item difficulty need to be "unidimensional" variables. For complex abilities, such as a student's ability in mathematics, the standard approach is to subdivide into a number of unidimensional variables. For example, the OECD's PISA study of mathematical literacy around the world reports mathematical literacy in "four separate one-dimensional models" (OECD, 2005, p. 191) of (i) space and shape, (ii) change and relationships, (iii) uncertainty and (iv) quantity, as well as providing an overall mathematics scale.

The aim of this paper is to provide an empirical examination of the assertions of many of our colleagues that a Rasch model would suit these data. However, as will be shown, the underlying assumptions of the model are not met. This case study provides an opportunity to consider fundamental questions about how learning proceeds. We propose that in this case, and presumably in many other cases, the better fitting model is one of *mapping learning* (i.e., recording students moving between categories of thinking) than of *measuring learning* (i.e., seeing how many questions they can answer correctly).

In the next section, the data set will be briefly described and then the initial results of a Rasch analysis will be reported. The following sections will demonstrate that the data contravene the fundamental requirements of the Rasch model, and then will explain why this is the case. The conclusion considers the general lessons about learning that arise from this situation.

The Data

The data to be analysed here were collected to study longitudinal and cross-sectional aspects of students' understanding of decimal notation. We were concerned with whether students can interpret a decimal number such as 0.456, for example, as 456 thousandths and as 4 tenths + 5 hundredths + 6 thousandths, and whether they know that it is slightly less than a half, more than 0.4, etcetera. Many results of the project have been published (e.g., Stacey, 2005; Steinle, 2004; Steinle & Stacey, 1998, 2003).

The data in this study are from a one page 30-item test designed to map students' understanding of the meaning of decimal notation. Every item on the test is of the same form. The instruction at the top of the page is *Circle the larger number in each pair of decimals*, and 30 pairs of decimals follow. As will be discussed below, item 6, for example, presents the pair {4.8, 4.63}[1] and students should circle 4.8 to be correct. Decimal comparison tests of this nature have been used extensively in mathematics education over many years, as well as in various large-scale testing programs. For example, item B10 in the category *Fractions and Number Sense* in TIMSS-R asked for the smallest number from this list {0.625, 0.25, 0.375, 0.5, 0.125}. The international average facility for Grade 8 students was 46%. As discussed below, our research showed that this is an over-estimate of the proportion of students able to order decimals because some have chosen the *correct answer* of 0.125 for the *wrong reason*.

The particular decimal comparison test used in this study is called DCT2. It was developed by the present authors, and Steinle (2004) gives the most complete account of its development and history and directions for its improvement.

DCT2 was completed by students from Year 5 to Year 10 (approximately 11 to 16 years of age) in 12 schools from a variety of socio-economic levels in Melbourne, Australia. Testing was administered to whole classes at

---

[1] Pairs are written 4.8/4.63 subsequently.

approximately 6 monthly intervals and many students will have been tested on multiple occasions, and their progress was tracked over time. The data set for this paper consists of 3531 test papers[2] of 30 dichotomous items. There were five test papers with every item completed incorrectly (i.e., scored 0 out of 30) and exactly 1200 test papers were completed with no errors (i.e., a score of 30 out of 30). Since these 1205 test papers do not contribute to the Rasch analyses, the effective size of the data set is 3531 – 1205 = 2326. Because many students were tested more than once (e.g., in Year 7 and also Year 8), the units of data are in fact the completed test papers, that is, students at given points of time, rather than the students themselves. For ease of reading however, and to align with the normal Rasch analysis language, we refer in this paper to the unit of analysis as the student.

Because the items all have exactly the same form and present the same task just with different numbers, it is *a priori* likely that they can be used to measure a developing unidimensional ability to order decimals. Simple evidence for this is that the percentage of students doing very well on DCT2 generally increases with year level, as shown in Table 1. The following sections will show, however, that the unidimensionality does not hold.

Table 1
Percentage of Students with High Scores (28, 29 or 30) by Year Level

	Year Level					
	5	6	7	8	9	10
Sample size	581	606	1254	457	350	283
High scores (28, 29)	10%	12%	13%	15%	14%	9%
Full marks (30)	9%	30%	42%	36%	44%	47%

Applying the Rasch model

Applying a simple logistic model (Rasch model) to the whole sample produced the estimates of item difficulty and fit statistics shown in Table 2 and the map of latent distributions and response model parameter estimates shown in Figure 1 (produced by Quest software, Adams & Khoo, 1993). The map in Figure 1 shows that the results are not ideal for Rasch analysis. Even with 1200 cases of perfect 30/30 results removed from the analysis, too many students find the items too easy (mainly above 0, a lot above 3), although Table 1 shows mastery is low. The estimates of item difficulties are bunched (mainly from -1 to +1 logit), with no items being identified as appropriately difficult or appropriately easy. In Figure 1, the items have been grouped horizontally into "types", which will be discussed later. At this stage it is sufficient to note that they are included in Table 2 and are sets of items which

[2] Steinle (2004) analysed 9862 test papers including these 3531 test papers, which were the earliest data collected and the only ones for which individual item data were entered electronically.

function similarly on decimal comparison tests.

Table 2

Estimates of Item Difficulty and Fit Statistics using the Rasch Simple Logistic Model on Whole Sample

Item number and comparison item		Item Type[a]	Estimate of item difficulty	Error	Weighted Fit (Infit)		Unweighted Fit (outfit)	
					MNSQ	*t*-values	MNSQ	*t*-values
1	0.4/0.457	S[b]	0.50	0.05	1.10	5.17	1.35	7.09
2	0.86/1.3	S[b]	-1.27	0.07	1.23	4.65	3.85	15.63
3	0.3/0.4	5	-0.64	0.06	0.81	-6.49	0.91	-1.18
4	1.85/1.84	5	-0.82	0.06	0.78	-6.74	0.65	-4.59
5	3.71/3.76	5	-0.77	0.06	0.77	-7.17	0.79	-2.67
6	4.8/4.63	1	0.38	0.05	1.15	7.67	1.05	1.06
7	0.5/0.36	1	0.54	0.05	1.20	10.44	1.14	3.16
8	0.75/0.8	1	0.57	0.05	1.17	8.85	1.07	1.57
9	0.37/0.216	1	0.81	0.05	1.30	15.04	1.33	7.32
10	3.92/3.4813	1	0.63	0.05	1.26	13.29	1.25	5.34
11	1.06/1.053	S[b]	0.51	0.05	1.22	11.22	1.14	2.93
12	4.08/4.7	3	-0.14	0.05	1.14	5.90	1.11	1.76
13	3.72/3.073	3	0.11	0.05	1.16	7.61	1.03	0.50
14	2.621/2.0687986	3	0.09	0.05	1.20	8.86	1.05	0.96
15	8.052573/8.514	3	0.01	0.05	1.16	6.96	0.96	-0.67
16	5.62/5.736	2	0.02	0.05	0.83	-8.70	0.83	-3.10
17	0.5/0.75	2	-0.04	0.05	0.83	-7.94	0.81	-3.51
18	0.426/0.3	2	0.05	0.05	0.80	-10.39	0.70	-6.18
19	2.516/2.8325	2	0.02	0.05	0.77	-11.79	0.66	-6.83
20	7.942/7.63	2	-0.03	0.05	0.76	-11.74	0.62	-7.85
21	4.4502/4.45	4	0.58	0.05	0.99	-0.72	0.99	-0.33
22	17.353/17.35	4	0.42	0.05	0.92	-4.32	0.87	-3.04
23	8.245/8.24563	4	0.55	0.05	0.98	-1.39	0.98	-0.47
24	3.2618/3.26	4	0.51	0.05	0.96	-2.02	0.93	-1.51
25	3.741/3.746	S[b]	-0.52	0.06	0.74	-9.73	0.72	-4.17
26	0.35/0.42	6	-0.62	0.06	0.75	-8.99	0.54	-7.20
27	2.186/2.954	6	-0.52	0.06	0.71	-11.14	0.51	-8.10
28	0.872/0.813	6	-0.51	0.06	0.71	-11.23	0.51	-8.23
29	0.038/0.04	S[b]	0.71	0.05	1.31	15.83	1.48	9.96
30	0.006/0.53	S[b]	-1.12	0.07	1.06	1.36	1.64	5.32

Note. [a]Item type is discussed later in the paper. [b]Some items do not belong to an item type and are referred to as Supplementary. MNSQ = Mean–square.

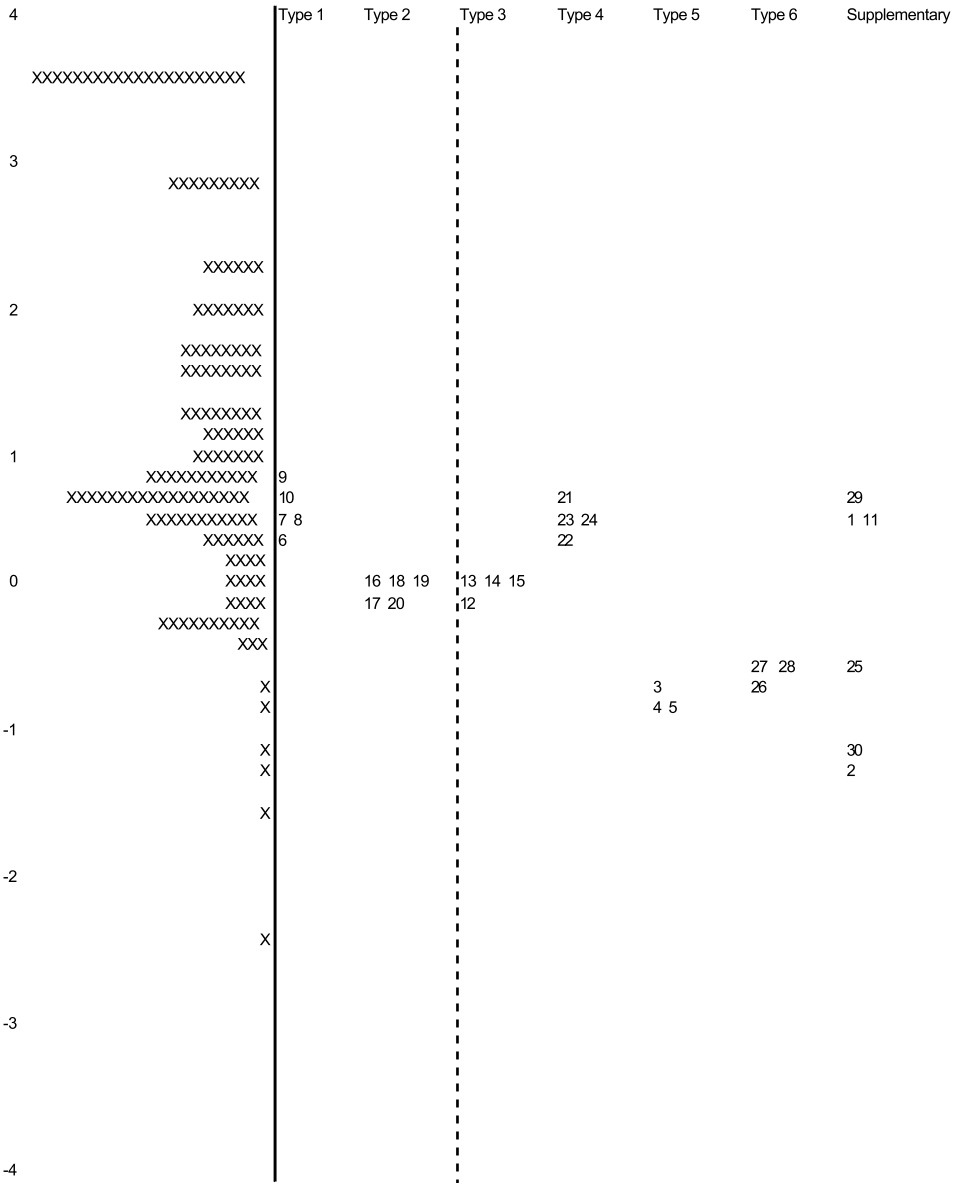

Figure 1. Map of estimates of individual ability and item difficulty .

Other Rasch indicators show that these data do not fit the model. Only a few of the 2 x 30 *t*-values for infit or outfit lie within the desirable range of (-2, 2). Given that many of the mean squares fit indicators are acceptable, one might conclude the misfitting *t* results are a consequence of the large sample size. Nevertheless, a less thoughtful approach to test construction than is ours would involve discarding items outside the desirable range, in this case leaving only 3 items in DCT2. A more sophisticated approach would examine the misfitting items carefully, to uncover the reasons behind the misfit. A helpful reviewer reminded us that Ben Wright often remarks that Rasch researchers take out stones from the soup and then study the stones: they do not just throw them away. In this case though, there are so few items with statistics within the desirable ranges, that all of the data from the modelling is suspect. The few items that supposedly 'fit' the model may not be good soup ingredients and the many items that do not fit may not be stones. Our conclusion in this paper will be that we are not making soup.

Differential Item Functioning by Year Level

Closer analysis reveals more deviations from the patterns expected for Rasch data. Figures 2 and 3 provide a plot of scores (vertical axis) against ability (horizontal axis) for Years 5, 7, and 9 for items 6 (4.8/4.63) and 16 (5.62/5.736), respectively, from Conquest software[3] (Wu, Adams, & Wilson, 1997). These two items have been selected as illustrative of other items in DCT2, as have the year levels 5, 7, and 9. The expected score curves, shown as a solid line, illustrate how the probability of being correct on an item is expected to increase as a logistic function of the ability of the students. Some plots, such as that for item 16 in Figure 3, show over-fit to the expected score curve. This item could be regarded as over discriminating: at all year levels good students do (slightly) better than expected, while low scoring students (below 0 logits) perform worse than expected on this item. Furthermore, middle scoring (0 - 1 logit) Year 5 students score better than expected on the item and *better than older students* of the same ability.

Figure 2 shows the actual score plots for years 5, 7, and 9 on item 6 (4.8/4.63), revealing a different pattern, which is exhibited by other items on DCT2 as well. At nearly all year levels (illustrated by Years 5 and 7 in Figure 2), the plots show that students of very low ability (e.g., less than 0 logits) have *higher* actual scores than students of medium ability (about 1 logit). This again indicates poor fit of the DCT2 data to the Rasch model.

---

[3] For convenience, output from both Quest and Conquest software is used in this paper. The very small differences in item difficulty estimates provided by the software algorithms do not affect the pattern of results.

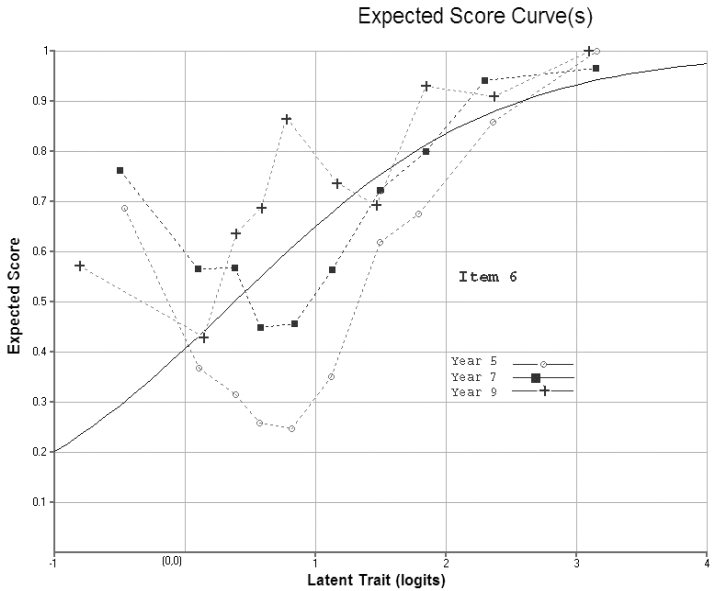

Figure 2. Actual score plots for Years 5, 7, and 9 and the expected score curve for item 6 (4.8/4.63).

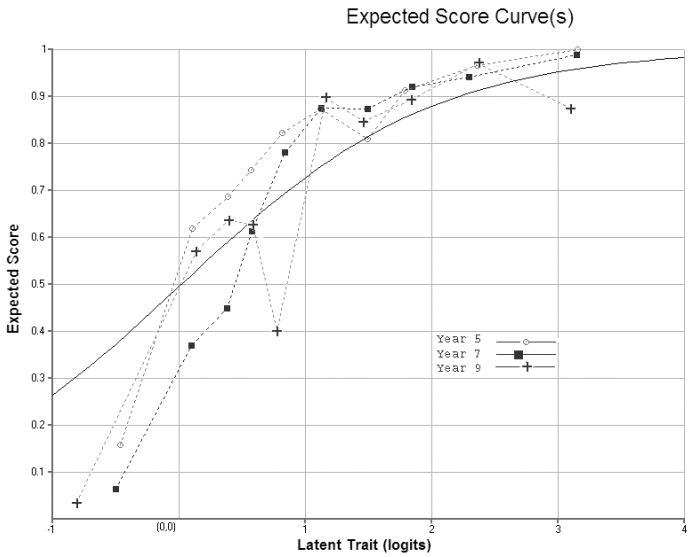

Figure 3. Actual score plots for Years 5, 7, and 9 and the expected score curve for item 16 (5.62/5.736).

To show this result more clearly, the estimates for item difficulty for each item type were averaged for each year level. These averages are presented graphically in Figure 4. This figure demonstrates that some item types are easier (i.e., have lower difficulty estimates) for Year 5 students than for Year 9 students, where classroom teachers might predict that younger students would be less able and hence not find items easier. Note that Year 7 is intermediate between Year 5 and Year 9 behaviour. Estimates of Year 6 are not graphed, as they are similar to those of Year 5; similarly, the estimates for Years 8 and 10 are like Year 9 values.

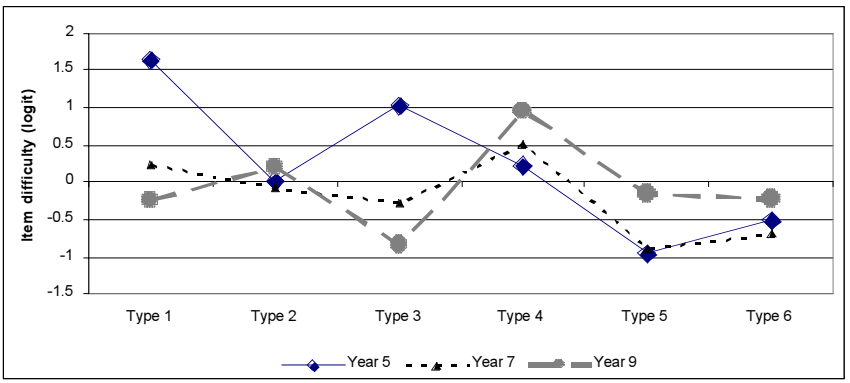

Figure 4. Estimates of item difficulty by year level averaged for each item type.

Differential Item Functioning Revealed Using Two Splitter Items

A second indicator of the poor fit of these data to the Rasch Model is obtained by splitting the sample on items 6 (4.8/4.63) and 16 (5.62/5.736). These items are of similar difficulty, with 64% of the non-perfect students correct on item 6 and 71% correct on item 16. Figure 5 shows the estimates of item difficulty obtained by Rasch analysis of the 29 remaining items (omitting item 6) and calculated on two samples – those getting item 6 correct and those with item 6 incorrect. Figure 6 shows similar information split on item 16. If the DCT2 data fitted the Rasch model's requirements for measurement, the set of students getting an item correct *should have higher ability* than the set of students getting the item incorrect, leading to the thicker line being consistently below the thinner line on both Figures 5 and 6. Inspection of these figures reveals that this is not the case with these data.

Furthermore, the estimates of item difficulty should be *ordered* similarly (allowing for error) for both sets, yet Figures 5 and 6 show that this is not the case. For example, consider the students who got item 16 correct (thick line in Figure 6). These students found items 17 to 28 *easier* than items 6 to 15. For the students getting item 16 incorrect, however, the reverse is the case; they found items 17 to 28 *harder* than items 6 to 15. Figure 5 shows a similar swap

of estimates of difficulty.

A comparison of Figures 5 and 6 is also revealing. Items 6 and 16 are of medium difficulty; the item difficulties of 0.38 (err: .05) and 0.20 (err: .05) (from Table 2) indicate that while item 6 is a marginally harder item, for our practical purposes they can be considered to be of equal difficulty. As such, the students who are correct on item 6 should be of *equal ability* to the students who were correct on item 16. A comparison of the plots for items 21 to 24 in Figures 5 and 6 reveal unexpected results. For students with item 6 correct, these items have a difficulty of approximately +1 logit compared with approximately -1 logits for the students with item 16 correct. In other words, the sub-sample with item 6 correct finds these items *more difficult* than the does the sub-sample with item 16 correct, even though we could infer them to be of equal ability overall. These differences show that there is not a uni-dimensional scale of difficulty underlying these items.[4]

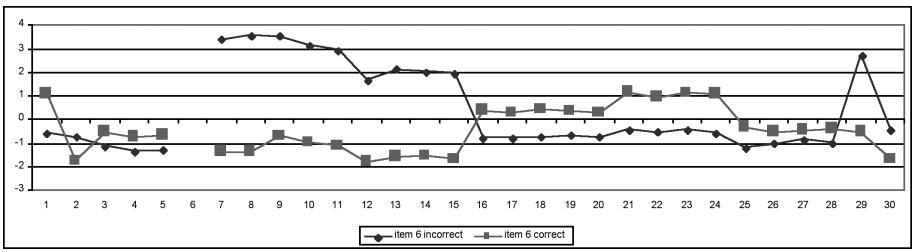

Figure 5. Item difficulty estimates, calculated on two samples having item 6 (4.8/4.63) correct and incorrect.

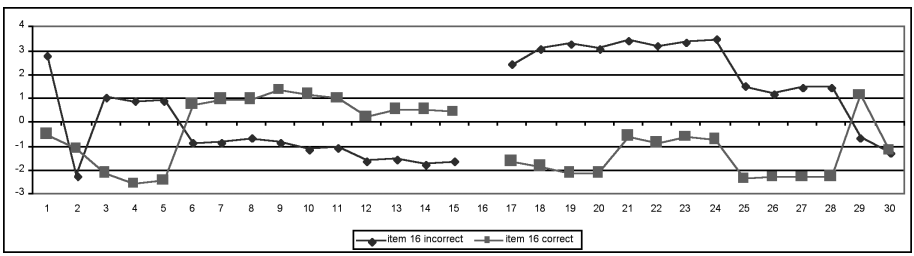

Figure 6. Item difficulty estimates, calculated on two samples having item 16 (5.62/5.736) correct and incorrect.

---

[4] The reader may deduce that a two-dimensional model may fit these data well: one dimension related to progress on items such as item 6 and one on progress with items such as item 16. However, these two item types do not describe the full picture; they have been used here to illustrate alternatives and are not comprehensive.

Why the DCT2 Test Framework is Inappropriate for Rasch Modelling

This section explains why the fit of these DCT2 data to the Rasch model is poor. The 30 items in DCT2 have been carefully selected to diagnose the way in which a student is interpreting or mis-interpreting decimal notation. As listed in Table 2, the 30 items have been classified on the basis of students' responses and mathematical analysis into 6 'item types' (with 6 unclassified supplementary items). A student's interpretation of decimal notation is diagnosed from their answers to the sets of items of each type. The item types are groups of items which *all* students are likely to regard as being equivalent questions. These types were constructed carefully, from an extensive knowledge of how students think about decimals. For each item type, the test designers aimed to select pairs of decimals which a given student would either get all correct or all incorrect, regardless of their ideas about decimal notation. Type 1, for example, consists of the five items 6 to 10. All of these items have equal whole number parts, decimal parts of unequal length, and no zeros in the decimal part, but the major feature is that the longer decimal is the smaller number. Items in type 2 are similarly defined, but the longer decimal is the larger number. Items in type 3 have one number with 0 in the tenths place, as well as meeting other criteria. If the type has been sufficiently well defined, any particular student's own interpretation of decimal notation applied consistently will result in that student getting all items in the type right or all wrong. An item type in our diagnostic test is therefore both a mathematical and psychological construction, and the definition of types depends on our knowledge of students' conceptions and misconceptions. This is in sharp contrast to the practice of those using Rasch measurement in achievement testing who attempt to choose or develop items in a hierarchy of difficulty to represent a single underlying ability dimension.

The item types were defined to elicit uniform responses from all students who had any of the 12 main misconceptions about decimal numbers that were known when DCT2 was created in 1997.[5] Their definitions depend on theoretical and empirical investigations of links between items by manual cross-tabs procedures (e.g., Stacey & Steinle, 1998) and by minimum message length cluster analysis (Nicholson et al., 2001). The clustering by type of the items in Table 2 and Figure 1 is further evidence that this has been quite successful. Careful analysis of the pattern of students' responses to each item type enables a high percentage of students to be classified according to the conceptions or misconceptions that they hold about decimal understanding. Details are presented in Steinle and Stacey (2003) and Steinle (2004), which also show how the proportion of students with each misconception varies with year level.

---

[5] More recent versions of the decimal comparison test, which diagnose further misconceptions especially for older students, are available from the authors.

Steinle and Stacey (1998) give a description of the 12 main misconceptions. Generally, younger students (e.g., in Year 5) bring their knowledge of whole numbers to their interpretation of decimals. Hence, they are likely to make errors on item 6 (4.8/4.63) and other Type 1 items, thinking that since $8 < 63$ then $4.8 < 4.63$. We label this, and several other variations, L behaviour (indicating *longer* decimals are likely to be judged to be larger numbers). This same thinking leads students to choose the answer correctly to item 16 (5.62/5.736) and other Type 2 items, but they are making *correct choices for the wrong reasons*. Thus, L students are incorrect on Type 1 items and correct on Type 2 items and their scores on the other types discriminate between varieties of L thinking[6].

A higher percentage of older students (e.g., in Year 9) tend to make the opposite choices, that is, a higher percentage of older students consistently select the decimals with fewer digits to be larger numbers, throughout the test. We label this as S behaviour (indicating *shorter* decimals are likely to be judged to be larger numbers). Some varieties of S behaviour are related to drawing false analogies with fractions or negative numbers, where larger whole numbers are associated with smaller numbers (e.g., $3 < 4$, but $1/3 > 1/4$ and $-3 > -4$). As with L behaviour, there are a variety of reasons for S behaviour, which can be tracked through item by item analysis and again, consistent thinking leads to some item types being answered correctly and some item types answered incorrectly. One group of S students considers all decimals with three places (i.e., thousandths) to be smaller than decimals with two places (i.e., hundredths), which are, similarly, smaller than decimals with one place (i.e., tenths). This group orders decimals of the same length as they would for whole numbers. Consistently applying this thinking leads students to be correct on item B10 in the category, *Fractions and Number Sense*, in TIMSS-R, which asked for the smallest number from this list {0.625, 0.25, 0.375, 0.5, 0.125}, but they would have been correct for the wrong reason. (They believe 0.5 and 0.25 are the largest two on the list.) Being correct on an item for the wrong reason characterises DCT2. It is one of the reasons why the DCT2 data do not fit the Rasch model, because these items break with the normal assumption that correctness on an item indicates an advance in knowledge (or ability) that will not be 'lost' as the student further advances.

What is evident from this discussion is that an individual decimal comparison item (or a group of comparison items of the same type) cannot be used to measure anything worthwhile. When students are correct on one item or on a set of items of the same type, it may be for the right reason or the wrong reason, or they may have guessed. Instead, it is the pattern of choices on all items (i.e. a student's responses to the test as a whole) that can be used to categorise a student's thinking reliably. As Bond and Fox (2001) commented

---

[6] For example, some L students will continue to make errors on Type 3 e.g., item 13 (3.72/3.073), whereas others will know that the 0 in the tenths column makes the second decimal small, perhaps as an isolated known fact, without more general place value knowledge.

(p. 197), "before we examine the fit of test items to the Rasch model, we should first examine the fit of the Rasch model to the test framework."

The deviations from expectations of the Rasch model demonstrated in Figures 2, 3 and 4 can be explained by knowledge of the major misconceptions about decimal notation that are most commonly held by students at different stages of schooling. Table 3 shows the distribution of L and S students from Year 5 and Year 9 (calculated by excluding the students with perfect or zero scores because they do not contribute to the Rasch estimates). Nearly half of the Year 5 students were classified as L, answering incorrectly on Type 1 and often on Type 3, and correctly on Type 2 and often also correctly on Types 4, 5 and 6 (depending on the variety of L misconception). This group was much less prevalent in Year 9, where a much larger contribution is made by the students classified as S (answering correctly on Type 1 and incorrectly on Type 2). Thus the item difficulty estimates by year level reflect the varying proportions of students holding the different misconceptions. Type 1 items are apparently *harder* for younger students (compared with older students), as they are more likely to hold an L misconception and therefore get these items wrong (see Figures 2 and 4). The exact opposite holds for Type 2 items (see Figures 3 and 4), which are judged *easier* for younger students (compared with older students). On this data set, the Rasch estimates of difficulty are lower for the items that are selected correctly by popular ways of thinking (correct or incorrect) at a given year level.

A finer grain analysis considering other item types can explain more of the detail of the Rasch estimates including many of the deviations from the expected score curves in Figures 3 and 4, but the picture remains the same. As students progress through school, learning more (e.g., learning about negative numbers), their ideas about decimal numbers change, and this may or may not result in more correct choices on this test. Again this is a break with the assumptions on which Rasch modelling is based. A student's total score on this test might increase or decrease depending on the particular misconception and the mix of items in the test. This does not fit the property of Rasch scaling stated in Swaminathan (1999), that "the number right score contains all the information regarding an examinee's proficiency level, that is, two examinees who have the same number correct score have the same proficiency level" (p. 49). Neither the total score approach of the classroom teacher, nor Rasch measurement estimates provides a felicitous summary of student performance on the decimal comparison items of the DCT2 test.

Table 3
Percent of Students without Perfect or Zero Scores classified as L or S, from Years 5, 7 & 9

Year	Sample size	L	S
5	531	46%	17%
7	731	23%	21%
9	197	11%	18%

Strong supporters of Rasch measurement might have several suggestions on how to make the data fit the model, overcoming the misfits that we have highlighted. For example, the differential item functioning revealed by the two splitter items might be overcome by using a two-dimensional model, so that competence in DCT2 is regarded as the outcome of two separate abilities. Splitting by items of other types is likely to reveal that further dimensions are required, and so a multi-dimensional model could be built. The differential behaviour of students by year level further suggests that different models for different year levels may be required. Eventually, the data might be made to fit the Rasch model quite well. However, our claim is not that Rasch modelling cannot be imposed on data about decimal conceptions, but rather that there is nothing to gain in following that approach in this and other cases. As the final section discusses, learning as revealed by answers to test items is not always of the type that is best regarded as 'measurable', but instead learning may be better mapped across a landscape of conceptions and misconceptions.

Mapping rather than Measuring Conceptual Change

This paper has been written as a response to the repeated suggestions made by colleagues over the years, which implied that we had been remiss in *not* using this Rasch analysis with our data. Surely we should take the opportunity to track growth by creating a *measure* of decimal understanding based on a proper scientific footing, as offered by Rasch analysis. This paper has demonstrated that these data do not fit the Rasch model. Nevertheless, while the test and the Rasch model are incompatible for very good reasons, our Rasch analysis of these DCT2 data provides very compelling empirical evidence based on our colleagues' own chosen analytical terms, that a measurement approach is not suited to the purposes of our diagnostic testing. Standard responses such as "eliminate items that do not fit the model" are inappropriate. Instead the lack of fit highlights some general lessons about the items, as has been discussed above, but also about learning and students' conceptual development. We set out to build a diagnostic device to detect particular misunderstandings about the decimal notation system, not to build a test of achievement. If teachers were to provide learning experiences more appropriate to children according to DCT2 diagnoses, we conclude that those children would fare better on the Rasch modelled tests so favoured by our colleagues in mathematics education.

Instead of testing a gradual accumulation of facts and skills, students who make errors on DCT2 reveal their conceptual understanding of decimals.[7] Understanding decimal notation may appear a very limited task,

---

[7] The test does not tap into the conceptual understanding of students who use an algorithm to compare decimals. Many of the 1200 students who scored 30 out of 30 would be in this situation. The strength of the test is to identify the way of thinking of students who make errors.

but full understanding requires mastery of a complex web of relationships between fundamental ideas. As students progress through school, they bring to bear different aspects of this complex web on their interpretation of decimals, and so the progress of learning in this domain by a large proportion of students can be characterised as movement between conceptions (and misconceptions), rather than steady progress "upwards" towards expertise. These misconceptions are of varying sophistication (e.g., some incorporate more complex place value ideas than others) but they are not ordered linearly from bad to good, as required by the Rasch model, with students getting more DCT2 items correct as they improve. Steinle (2004) investigated the possibility of establishing a hierarchy of misconceptions and came to the conclusion that it was best based on "readiness to move to expertise". In Steinle's sense, a better misconception to have is defined as one with a higher probability of becoming an expert the next time the test is administered. This hierarchy does not show up in the total scores on the test or the Rasch ability estimates for students. Instead a student's score on DCT2 and the Rasch ability estimate is an artefact of two things: (i) the relative proportion of items from each item type on the test (because students who are correct on the more frequently occurring item types get higher scores); and (ii) the relative proportion of students with given ways of thinking about decimals in the sample. This latter factor is in turn affected by underlying progress that students make, but also sample characteristics including the age profile.

Had the data been described well by the Rasch model, it would have implied that students begin by understanding some 'simple' decimals and hence develop the ability to compare them, and then gradually learn about decimals of more complex types and develop the ability to compare those too. A paradox is that although this is not a reasonable description of *students' thinking* and its development, it does seem a reasonable description of the *teaching process*. Students begin to learn about decimal numbers with one place (e.g., 0.3 and 0.4), and learn to interpret them as 3 tenths and 4 tenths so that they can see that $0.3 < 0.4$, and so forth, and to add and subtract them. Next they learn about decimals with two decimal places, then three and more, and also learn the 'reunitising' that links decimals with different numbers of decimal places (e.g., that $0.3 = 0.30$), which is needed for addition and subtraction of ragged decimals, and so on. Later, they learn about longer decimals and more complex operations. In this sense they gradually build up knowledge. Items that tested a wide range of such facts and skills would very likely be able to form an acceptable uni-dimensional Rasch scale—but it would be a far different test.

As students change their ideas about decimals, they make different choices on the DCT2 items. Sometimes these later choices are correct on items that were previously incorrect, and sometimes these later choices are incorrect on items that were previously correct. We propose that changes in students' ideas are best thought of as movement between categories (with

more students having reached the goal as they get older) rather than as movement up a scale. The best model of learning in this case is as 'around a landscape' rather than 'up a ladder'.

A teacher looking at the results of some of her students on DCT2 commented to us that they had "just a few more little things to learn." In fact her students had very little idea about place value: they thought that the whole number indicates the number of dollars and the first two decimal places indicates the whole number of cents. This interpretation happens to produce correct answers on most item types on DCT2. It appeared to the teacher that there was only one item type these students needed to learn about (Type 4 in Table 2 e.g. 4.4502/4.45), but in reality they needed a major change in their understanding. Their conception incorporated no understanding of place value beyond isolated information that the decimal point separated two whole numbers (the number of dollars from the number of cents) and the later digits did not matter. To the teacher, thinking of learning as 'accumulating facts' and impressed by the score on DCT2, the students knew a lot and were well on the way to mastery, but from the point of view of learning principles and making connections between ideas, these students had hardly begun.

This paper is not intended to imply that the Rasch model cannot describe progress in learning in many situations very well. It is very well suited to descriptions of students' learning in domains of knowledge which are viewed from a reasonable distance, so that some overall progress of 'learning more' is evident, perhaps over some years of schooling or with tasks clearly increasing in scope and complexity. We have provided evidence, however, that not all learning is well described like this. By setting out the assumptions behind the construction of a scale for measurement, Rasch theory has helped us see that that different aspects of learning need to be tracked with fundamentally different tools. There will surely be many other examples beyond the case of decimals, where aspects of learning need to be mapped rather than measured.

Acknowledgments

We wish to thank Andrew Stephanou of the Australian Council for Educational Research, who carried out initial Rasch analysis on the data, and especially Nathan Zoanetti of the Assessment Research Centre, University of Melbourne, who conducted the final analyses and assisted in the interpretation.

References

Adams, R. J., & Khoo, S. T. (1993). Quest: The Interactive Test Analysis System. [Computer software]. Melbourne: Australian Council for Educational Research.

Bond, T. G., & Fox, C. M. (2001). *Applying the Rasch Model: Fundamental measurement in the human sciences*. Mahwah, NJ: Lawrence Erlbaum.

Nicholson, A., Boneh ,T., Wilkin, T., Stacey, K., Sonenberg, E., & Steinle, V. (2001). A case study in knowledge discovery and elicitation in an intelligent tutoring application. In J. Breese & D. Koller (Eds.), *Proceedings of the 17th conference on Uncertainty in Artificial Intelligence* (pp. 386—394). San Francisco: Morgan Kaufmann.

Programme for International Student Assessment (PISA). (2005). *PISA 2003 technical report*. Paris: Organisation for Economic Co-operation and Development.

Stacey, K., & Steinle, V. (1998) Refining the classification of students' interpretations of decimal notation. *Hiroshima Journal of Mathematics Education. 6* , 49—69.

Stacey, K. (2005). Travelling the road to expertise: A longitudinal study of learning. In H. L. Chick & J. L. Vincent (Eds.) *Proceedings of the 29th annual conference of the International Group for the Psychology of Mathematics Education* (Vol. 1, pp. 19—36). Melbourne: PME.

Steinle, V. (2004). *Changes with age in students' misconceptions of decimal numbers*. Unpublished PhD thesis, University of Melbourne, Australia.

Steinle, V., & Stacey, K. (1998). The incidence of misconceptions of decimal notation amongst students in grades 5 to 10. In C. Kanes, M. Goos, & E. Warren (Eds.), *Teaching mathematics in new times.* (Proceedings of the 21st annual conference of the Mathematics Education Research Group of Australasia, Vol. 2, pp. 548—555). Gold Coast, QLD: MERGA.

Steinle, V., & Stacey, K. (2003). Grade-related trends in the prevalence and persistence of decimal misconceptions. In N. A. Pateman, B. J. Dougherty & J. Zilliox (Eds.), *Proceedings of the 27th annual conference of the International Group for the Psychology of Mathematics Education* (Vol. 4, pp. 259—266). Honolulu: PME.

Swaminathan, H. (1999). Latent trait measurement models. In G. N. Masters, & J. P. Reeves (Eds.), *Advances in measurement in educational research and assessment* (pp. 43—54). Oxford: Pergamon.

TIMSS-R. *IEA's Repeat of the Third International Mathematics and Science Study at the Eighth Grade.* Retrieved August 30, 2005, from http://timss.bc.edu/timss1999i/pdf/t99math_items.pdf

Wu, M. L., Adams, R. J., & Wilson, M. R. (1997). ConQuest: Multi-Aspect Test Software [Computer software]. Melbourne: Australian Council for Educational Research.

Authors

Professor Kaye Stacey, Faculty of Education - Science and Mathematics Education, University of Melbourne VIC 3010. E-mail: < k.stacey@unimelb.edu.au

Vicki Steinle, Faculty of Education - Science and Mathematics Education, University of Melbourne VIC 3010. E-mail: < v.steinle@unimelb.edu.au